A fast and economic ontology engineering approach towards improving capability matching: Application to an online engineering collaborative platform

Xiao Ma[*a], Jay Bal [a], Ahmad Issa [a]

[a] University of Warwick, WMG, International Institute for Product and Service Innovation(IIPSI), Coventry, CV4 7AL, United Kingdom

[*] Corresponding Author, Tel. +44 2476 524718;
E-mail addresses: x.ma@warwick.ac.uk (Xiao Ma), jay.bal@warwick.ac.uk (Jay Bal), a.issa@warwick.ac.uk (Ahmad Issa).

Abstract

Fulfilling needs through internal and external resources is a key business requirement. To better enable this, description of both needs and resources, using a common domain language is required. Using techniques from Social Network Analysis (SNA) this paper describes a SENSUS-based methodology which generates domain ontology that can provide the breadth and depth of coverage required for automated need and resource matching systems. The mechanism described also enriches the semantic relationships in the generated ontology to form a network structure. This enables concept investigation to be undertaken from multiple perspectives, with fuzzy matching and enhanced reasoning through directional weight-specified relationships. The methodology was used to derive an ontology for engineering and tested against a traditionally derived and structured ontology. The methodology has the flexibility and utility to be of benefit in a wide range need and resource matching business applications.

Keywords: Information Systems, collaboration platforms, Knowledge Engineering, Ontology Engineering, virtual collaboration, Virtual Team, Semantic Web, Semantic Network, Interoperability, Social Network Analysis.

1 Introduction

In this paper we describe a new and novel way to automatically generate ontology that can be used in information systems to reason and structure information. The application of the methodology is illustrated in an engineering sector case study, The West Midlands Collaborative Commerce Marketplace. This online portal helps match tender opportunities with companies that have the right capability, and can help form supply chains or consortia with all the capabilities required to enable collaboration to exploit an otherwise very difficult to address opportunity. WMCCM is a representative of a generalizable collaborative platform [1] or virtual organization [2]. Many other "matching" type platforms exist in many sectors, ranging from personal "dating" to business "sourcing".

An ontology is a formal representation of knowledge as a set of concepts within a domain, and the relationships between those concepts. In information science it is used to reason about the entities within that domain, and may be used to describe the domain. The increasing need for information exchange within and between market sectors has driven the interest in ontology generation [3, 4]. Ontology are increasingly used in knowledge management systems, medical and bio-informatics and play a key role in the semantic web and grid computing. Engineering was among the earliest sectors to benefit from ontology, and ontology in this sector are considered to be more mature than in others.

Engineering ontology are structured and populated to fit their special needs. Thus the way they are intended to be used determines how they are formed. Application orientation is also emphasized in the Developing Ontology-Grounded Methods and Applications (DOGMA) approach [5], where the ontology structure is designed as "double articulation" – a domain specific articulation and an application specific articulation. The practical requirement of the ontology application environment also drives the engineering ontology discussed to stretch the traditional ontology boundaries in terms of representation and weight specification.

Several ontology have been built by various organisations in the engineering sector, often in the form of industrial classifications to allow information exchange among organisations, such as United Nations Standard Products and Services Code[2] (UNSPSC) and UK Standard Industrial Classification (SIC) [6]. However, these classifications showed a lack of broad cover of the classes, especially with regard to the actual products and services, and insufficient relationships to demonstrate inheritance and commonality among classes [7]. In addition, the condensed classes produced by experts did not have enough attributive descriptions around concepts. In other words, there were a small number of words to cover a much larger generic keywords variation in natural language information. Finally, classes (concepts) proposed by such sources tended to stay at a higher level compared with the company/user proposed classes. The high level classes were found not to be specific or detailed enough to differentiate between the competences proposed by companies. These issues suggest that directly summarising ontology from existing sources (a single top-down procedure) may not satisfy the practical requirement of the collaboration platform for broad coverage and rich internal relationship.

[2] http://www.unspsc.org/

The paper first investigates current methods for ontology generation and then goes on from an analysis of their shortfall for industrial applications, to describe a new methodology which addresses some of these key issues.

2 Related Work

A review of ontology engineering methodologies, including Cyc Base [8], TOVE [9], On-To-Knowledge [10], METHONTOLOGY [11] and SENSUS [12], was conducted to assess their applicability to the notion of "economic, quick and reliable" ontology generation (Appendix B). Namely in this research, these criteria refer to a requirement for little or no reliance on domain experts, (fast) speed of corpus building and corpus structure analysis, and applicability to multiple (or cross) domains. The various ontology engineering methodologies were also evaluated on their coverage of the domain and the richness of the internal relationships.

Cyc methodology was applied to build the Cyc Knowledge Base, which is one of the top level ontology that SENSUS refers to. It was constituted by manually adding over a million pieces of consensus knowledge statements. Domain experts were the starting point for building the knowledge base. Most of the knowledge in the system would be based on the opinions of a group of experts. However this may not be sufficient to cover wider perspectives in the field and the common vocabulary of non-professionals. Domain experts were also needed in all of the later stages, resulting in a costly way of building such ontology.

TOVE's approach proposed a methodology in a linear process with detailed techniques at each stage. However, the technique details limited the methodology into wider application environment. For instance, using "first order logic" to specify the terms and relationships led to its inapplicability for developing ontology, which requires other types of binary relationship, i.e. semantic relationship. Although this relationship could be altered, it was bounded to TOVE's development environment, and any alterations might require much greater consideration so as to modify the remaining part of the methodology, for use in other projects.

On-To-Knowledge and KACTUS improved the linear process by suggesting a development cycle in order to enable knowledge reuse and continuous improvement (even for application in different domains). Researchers [11, 13-15] have integrated formalised methodologies with ontology reuse methods, such as METHONTOLOGY. Despite a relatively comprehensive methodology with detailed techniques in ontology engineering, METHONTOLOGY did not appear to have the flexibility to rapidly respond to changes within the domain due to its manual corpus construction processes.

The proposed methodology for building the ontology is based on the principle that the ontology building should be initialised by linking specified keywords to the target source. SENSUS [12] constructs ontology for a domain from the foundation of a large knowledge base, or ideally, a previous large ontology. However, it does not engage in a traditional reusing or re-engineering process. It identifies key domain specific terms, a.k.a. seeding words, and then links them to the large ontology. Afterwards, the terms irrelevant to the new ontology can be pruned from the large source ontology. The processes undertaken in the SENSUS approach are shown in Figure 1.

Figure 1: SENSUS Approach to developing ontology

This approach contains unique characteristics that provide advantages over the other methodologies:

- It is an obvious improvement that SENSUS does not require constant input from domain experts: it only needs the initial seeding terms and their relationships to the knowledge base.
- SENSUS combines corpus construction and ontological analysis in one process, unlike others [8, 11]. SENSUS thus ensures the terms collected are semantically connected to the seeding terms.
- SENSUS can act like a shared foundation to allow other ontology to be connected together and to share their terminology and relationships [12].
- Extracting related terms from the same sources through different seeding words is similar to seeing the same knowledge from different perspectives. This in theory could result in fuzziness around any given concept depending on the number of perspectives chosen. Thus the SENSUS ontology construction method may be capable of building cross-domain ontology.

Despite these benefits, it is difficult to apply SENSUS directly for our need: resource matching requirement, as there is insufficient detail on the techniques suggested to apply it. In addition, SENSUS did not propose any post-development stage, a development life cycle or project management mechanism which would help in industrial applications. Therefore, this research used the SENSUS approach as a foundation approach and developed techniques to formulate a new methodology that met the needs for quick, economical, reliable, and multi-domain ontology construction.

3 Research Methodology

The SENSUS methodology recommended that the ontology building should be initialised by linking specified keywords to the target source.

3.1 Data Source Selection

Word clustering is a technique for partitioning the words that describe a domain into subsets of semantically similar words and is important in a number of Natural Language Processing tasks. The sets of words that describe the domain will be called ‘keywords sets’ hereafter. There are basically two main data sources (corpus) that could be used to generate these keywords:

1. Directly collected expert and user data: first hand data;

2. Directly reused or extracted data from existing data sources which contain words with either their semantically similar or semantically related relationships. There are five types of such sources:
    a. Thesauri or Dictionaries: representing a dictionary type;
    b. WordNet: representing general lexical ontology or databases;
    c. Industry/Government Codes;
    d. Ontology search engines, such as OntoSearch, OntoSelect and Swoogle, which represent searchable ontology databases that index lists or directories of ontology
    e. Search Engine Indexes: Current use of search engine indices in ontology engineering mainly concentrates on refining current ontology towards including fuzzy logic [16] and new knowledge acquisition[17, 18].

The research requirement for less reliance on domain experts, broad coverage of concepts and rich internal relationships means that the use of first hand data is not suitable since it requires significant input from domain experts. In addition, the use of semantic relatedness means that thesauri/dictionaries and WordNet are not suitable source knowledge bases. Thus a general search engine index, which crawls all types of web pages on the Internet, may better suit the need of this research for a broad coverage incorporating the latest developments and rich relationships among the terms.

There are many popular search engines available across the Internet, such as Google and Bing. Among these, Google has been widely regarded as the leader with the indexed content and popularity [19, 20]. Uniquely, Google provides a method – Google Sets [21] – to generate "on-topic" terms based on given examples. This method seems to provide an opportunity to generate domain related terms with wider but not chaotic relationships. It also allowed users to query via a standard HTTP GET command, so that queries can be automated via computer programs for various keywords or parameter settings. (Google Set was discontinued in 2011 as a separate tool; it is now available as an "autofill" function[3] in Google Spreadsheet.)

3.2 Seeding Word Configuration

Google Sets is a word clustering tool which extracts semantically associated words from the Google index. In our case it could link initial domain seed words to the Google index via their semantic relationships. This allows us to extract all of the semantically related words to the initial seeding words from the Google index. In other words, any seeding word should be connected to its semantically associated terms in the index source. These terms are group(s) of concepts representing similar domain concepts to the seeding words. Since Google Sets is a word clustering tool which extracts semantically associated words from the Google index, this was used to link the initial seeding words to the Google index via semantic-relatedness relationships. Google Sets has several parameters that can be altered through settings, and a study of these parameters was conducted so that they could be configured to provide the best results.

Early experiments to test the quantity and quality of predictions showed that paired keywords generated better results than any other option. Paired seeding words had the advantage of producing more focused domain terms, and it seems that paired seeding words particularly

[3] https://support.google.com/drive/answer/75509?hl=en

benefitted the domain description density for both less focused domains and more naturally focused domains. Therefore paired seeding words were utilised for generating the engineering ontology.

However, a further issue was the need to avoid seeding words that had high potential for misleading the search direction. Therefore, further experiments were conducted to identify the minimum number of seeding word pairs required to provide reasonable fault tolerance. The results showed that two pairs of keywords appear to be the minimum required. However, two pairs of seeding words may produce predictions around two subject areas. In an extreme case (Figure 2), if a pair did not produce any target domain prediction at all, the experiment may end up with two separate distributions of terms, with no overlap. In such a case, the resulting corpus of terms may not target any particular domain, and further expert guidance may be required. Using three pairs, the system will better tolerate poor seeding word choices, and ensure the output is more reliable.

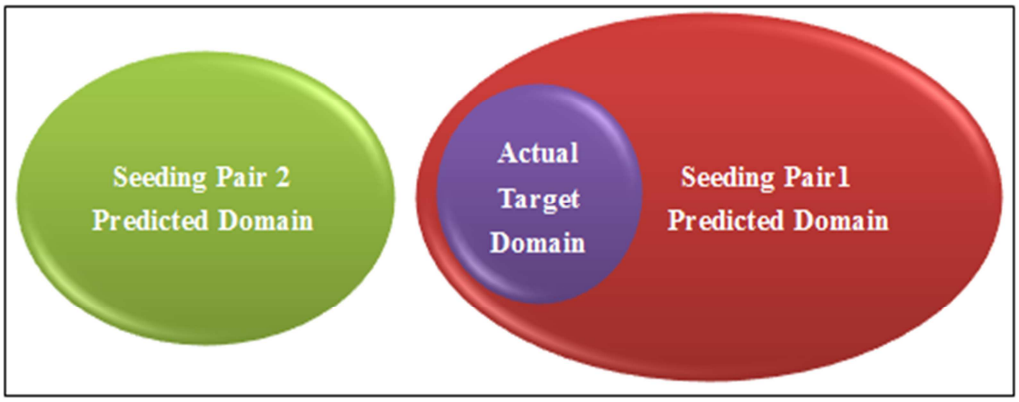

Figure 2: Complete Prediction Separation of Two Pairs of Seeding Words

3.3 Seeding words Selection

A Delphi approach to collect seeding words for a subject area from domain experts was adopted [22]. This method collects the opinions of different individuals in order to increase the opportunity of picking objective seeding words and minimize subjective bias from direct study of the application environment.

3.4 Corpus Construction

Google Sets was used to generate semantically related terms from the initial seeding word pairs. However, the resulting terms were too few to represent any practical domain or to yield any statistically relevant results. To generate more keywords, the resulting terms were reprocessed as new seeding word pairs to obtain more predicted terms. After this second round of seeding there was better coverage of the domain, but still insufficient concepts and relationships to yield any statistical reliability. Therefore the terms generated from the second round were used as seeding words to derive third level predictions.

This method is known as “Snowball Sampling” and is common in social studies and statistics, especially within social network analysis [23]. This approach generates a large collection of related entities to construct complex social networks [24]. There are associated social network analysis techniques to uncover more facts about such a network. The methodology may be represented by the following formula.

$$S_{(x,y)} = f_{GS}(x,y) = \left\{k_1^{x,y}, k_2^{x,y}, \ldots, k_{n_{x,y-1}}^{x,y}, k_{n_{x,y}}^{x,y}\right\}$$

Function $f_{GS}(x,y)$ is the process to capture Google Sets results by using given paired seeding keywords x and y. Set $S_{(x,y)}$ is the collection of keywords from $k_1^{x,y}$ to $k_{n_{x,y}}^{x,y}$ generated from seeding keyword pair x and y. The multiple iterations of this formula in achieving the "snowballing" effect are explored in more detail in Appendix A.

This would derive a large collection of possible terms for the ontological corpus, which would subsequently serve as primary data for further network analysis and the ontological structure pruning. The network structure may result in different facets of the ontology to suit different applications in different domains. Therefore, this paper will illustrate the ontological analysis of the primary data within the context of an online collaborative platform, the West Midlands Collaborative Commerce Market Place (WMCCM).

4 The Collaborative Platform Case Study

WMCCM is a web collaboration platform matching "need" in the form of requests for "competence" and resources originating from tenders, with the capabilities of small and medium size businesses (SMEs). Often the overall tender needs can only be met by enabling collaborations among independent businesses to combine their individual competencies [25]. WMCCM currently has over 13,000 member companies, deals with over 60,000 tenders per year and each year these companies win over £4bn worth of tenders. In order to automate the matching process between companies and tenders, WMCCM classifies company competencies against a three level ontology (figure 3). It also semantically analyses every incoming tender to identify what competencies are required and maps these onto the same ontology through a naive Bayesian classifier. This allows the WMCCM system to forward tenders to companies that have the right capability, and to support the creation of new partnerships. An example tender is shown in figure 4.

A key factor affecting the effectiveness of the matching functions is the quality of the ontology that links tenders with company capability and competency. The WMCCM engineering ontology was built in an orthodox way by the re-use of previously published ontology and adaptation/modification by experts. Thus it followed a mixed approach: lower levels were derived from actual company interview information; upper levels from standard classifications such as the United Nations Standard Products and Services Code (UNSPSC) and United Kingdom Standard Industry Classification (SIC).

UNSPSC was designed as an upper level ontology to facilitate business for quicker and more accurate procurement, marketing and sales. It was designed for high level guidance, but it does require adaptation to be practical at the country and region level [26]. The UK SIC is one adaption of UNSPSC and is the standard industrial classification widely used in the UK. It is used to categorise businesses in accordance with the scope of their economic activity [6]. Although fundamentally UNSPSC and SIC were supposed to represent the same knowledge and its structure, UNSPSC lacks domain coverage in some areas and depth in others. With regard to the actual products and services, described by the ontology there are insufficient relationships to provide

inheritance and commonality among classes [7]. This illustrates that while many ontology used in industrial applications have reused such sources, they still require considerable consultancy from domain experts to clarify the relationships between such sources. [27].

Directly applying ontology from existing sources does not satisfy the requirement for a broad coverage ontology to address the contextual variety of tenders available on WMCCM and the provision of rich internal relationships, with connecting terms to address domain specialisms. Thus like many other systems, WMCCM used a mixture of top down derivation and bottom up synthesis to derive its ontology, enabled by collecting terms and relationships from actual business users to supplement the top level SIC orientated coding.

However, this customisation still did not fully meet the needs on WMCCM's tender matching process. The source ontology (based on UNSPSC and SIC) lacked the necessary level of fuzziness/redundancy to work effectively with human oriented systems, such as the text based tenders needing analysis and interpretation in WMCCM. The reuse of high level ontology only provides the guide structure and description of the domain knowledge, but lacks relationships to terms that are not strictly bounded by the core domain terms, but that are necessary for fuzzy interpretation. The required fuzziness may be achieved by increasing the number of semantic relationships identified that are not exclusive to that particular engineering domain. In order to address these issues, the aforementioned methodology was used to derive engineering ontology quickly, economically and reliably and that meets the needs of WMCCM system.

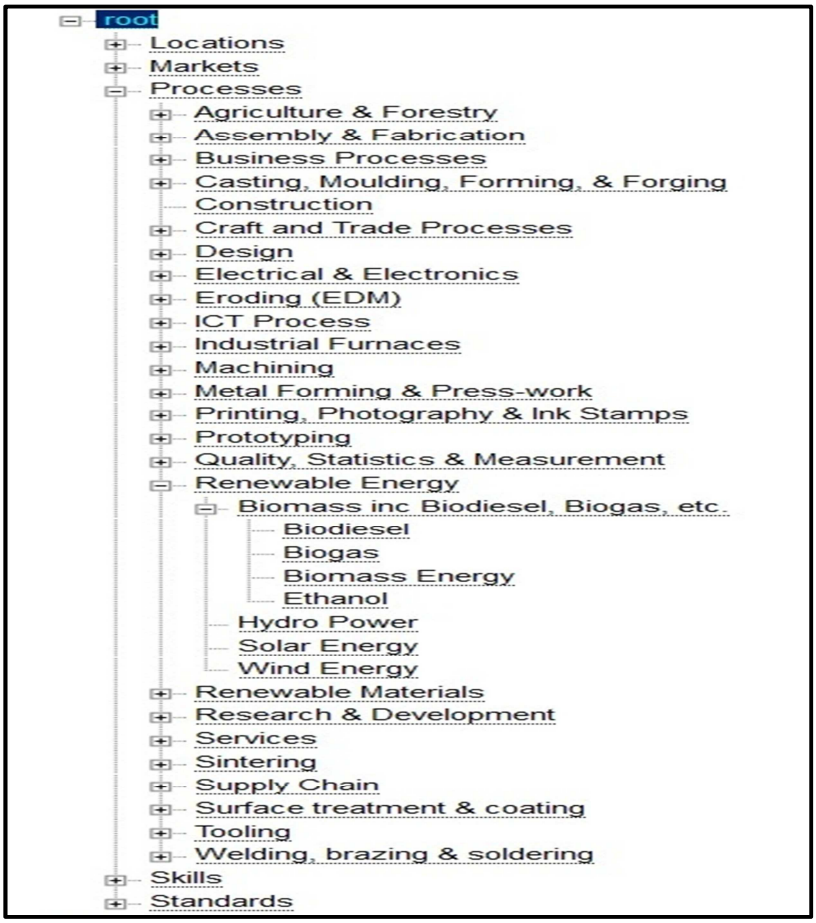

Figure 3: Current WMCCM Ontology[4]

[4] It is a three level tree structure, where only the "Renewable Energy" and "Surface Treatment & Coating" sections are expanded in this figure.

United Kingdom-Bristol: Furniture (incl. office furniture), furnishings, domestic appliances (excl. lighting) and cleaning products

Tender Closed - contact the successful bidder for subcontract opportunities	Name and address of economic operator in favour of whom the contract award decision has been taken Workscape LTD Unit 1, Westpoint Business Park SN14 6RB Chippenham UNITED KINGDOM
Direct Contact	Address: Email: Fax: Other:.
Type of Document	7 - Contract award
Title	United Kingdom-Bristol: Furniture (incl. office furniture), furnishings, domestic appliances (excl. lighting) and cleaning products
Publication Date	20140109
Number of Document	7149-2014
Number of the Official Journal	6/2014
Date sent to EUR-OP	20140107
Heading	01506
Nature of Contact	2 - Supply contract
Type of Procedure	4 - Negotiated procedure
Regulation of Procurement	4 - European Union
Type of Awarding Authority	4 - Utilities
Type of Bid Required	9 - Not applicable
Awarding Criteria	2 - The most economic tender
CPV Product Code	39000000 39153000
CPV Product Name	Furniture (incl. office furniture), furnishings, domestic appliances (excl. lighting) and cleaning products Conference-room furniture
Country Code	UK
Original Language	EN
Name of Awarding Authority	BRISTOL WATER PLC
Town of Awarding Authority	BRISTOL
Abstract - detail of work to be done	Bristol Water Furniture Broker as part of the Bristol Water Head Office Refurbishment Project. The successful company will be responsible for offering a complete project management and supply chain management solution to enable the successful procurement of Bristol Water furniture for the entirety of the Bristol Water Head Office Refurbishment Project in order to achieve the best possible value for money, within an agreed budget and timescales. It is anticipated that this project wil have three sub-phases due to the requirement to keep the vast majority of existing head office building staff on site during the works. The programme for furniture installation is February 2014, April 2014, October 2014 and April 2015. CPV: 39000000, 39153000.
Document Text	Contract award notice – utilities Directive 2004/17/EC Section I: Contracting entity I.1) Name, addresses and contact point(s) Bristol Water PLC PO Box 218, Bridgwater Road, Bedminster Down Contact point(s): Procurement Contracts Manager For the attention of: Zac Coley BS99 7AU Bristol UNITED KINGDOM Telephone: +44 1179341214 E-mail: contracts@bristolwater.co.uk Internet address(es): General address of the contracting entity: http://www.bristolwater.co.uk/ I.2) Main activity Water I.3) Contract award on behalf of other contracting entities The contracting entity is purchasing on behalf of other contracting entities: no Section II: Object of the contract II.1) Description II.1.1) Title attributed to the contract Bristol Water Furniture Broker as part of the Bristol Water Head Office Refurbishment Project. II.1.2) Type of contract and location of works, place of delivery or of performance Supplies Purchase II.1.4) Short description of the contract or purchase(s): The successful company will be responsible for offering a complete project management and supply chain management solution to enable the successful procurement of Bristol Water furniture for the entirety of the Bristol Water Head Office Refurbishment Project in order to achieve the best possible value for money, within an agreed budget and timescales. It is anticipated that this project wil have three sub-phases due to the requirement to keep the vast majority of existing head office building staff on site during the works. The programme for furniture installation is February 2014, April 2014, October 2014 and April 2015. II.1.5) Common procurement vocabulary (CPV) 39000000, 39153000 II.1.6) Information about Government Procurement Agreement (GPA) The contract is covered by the Government Procurement Agreement (GPA): no II.2.1) Total final value of contract(s) Section IV: Procedure IV.1) Type of procedure IV.1.1) Type of procedure Negotiated with a call for competition IV.2) Award criteria IV.2.1) Award criteria The most economically advantageous tender IV.2.2) Information about electronic auction An electronic auction will be used: no IV.3) Administrative information IV.3.2) Previous publication(s) concerning the same contract Contract notice Notice number in the OJEU: 2013/S 184-318406 of 21.9.2013 Section V: Award of contract V.1) Award and contract value Contract No: RFX-00000006 Lot title: Bristol Water Furniture Broker as part of the Bristol Water Head Office Refurbishment Project V.1.1) Date of contract award decision: 17.12.2013 V.1.2) Information about offers Number of offers received: 8 Number of offers received by electronic means: 1 V.1.3) Name and address of economic operator in favour of whom the contract award decision has been taken Workscape LTD Unit 1, Westpoint Business Park SN14 6RB Chippenham UNITED KINGDOM V.1.4) Information on value of contract Initial estimated total value of the contract: Value: 700 000 GBP Excluding VAT Total final value of the contract: Value: 700 000 GBP Excluding VAT V.1.5) Information about subcontracting The contract is likely to be sub-contracted: no Section VI: Complementary information VI.1) Information about European Union funds The contract is related to a project and/or programme financed by European Union funds: no VI.3) Procedures for appeal VI.3.1) Body responsible for appeal procedures Body responsible for mediation procedures Bristol Water PLC PO Box 218, Bridgwater Road, Bedminster Down BS997AU Bristol UNITED KINGDOM E-mail: contracts@bristolwater.co.uk Telephone: +44 1179341214 VI.4) Date of dispatch of this notice: 7.1.2014
Processes:	Installation, Procurement, Supply Chain, Supply Chain Management
Skills:	Supply Chain Management
Markets:	Domestic Appliances, Furniture, Water
Originator:	Tenders Electronic Daily
WMCCM Contact Name:	See tender for contact point, please login, you may need to register first.
WMCCM Contact Phone:	

Figure 4: An example of WMCCM tenders after it has been analysed and formatted by the ontology system

4.1 Primary Data

Three pairs of initial seeding words (drilling & cutting, milling & sawing, and turning & grinding) to represent the "machining" domain were obtained from the WMCCM project team. From these, 10,660 unique terms with 266,176 relationships among them were automatically generated using Google Sets and the procedure described in section three. Previously, WMCCM had used traditional manual processes to collect 862 unique concepts with 2,126 relationships from both the SIC and domain experts. The new ontology contained fifty times more terms, and more than a hundred times the number of internal relationships compared with the original WMCCM ontology.

These terms and their relationships formed a "concept" network of terms. This network is similar to many social networks and there are well established social network analysis methods which can be applied to the collected data to conduct ontological analysis.

4.2 Network analysis of the ontological structure

Analysis of the ontological structure reflected the later stage in the development cycle (figure 5): finding the "roots" – representatives of the network; clarifying links between new domain terms and "roots"; clustering sub-trees and defining the boundaries of subtrees and of the whole network. This analysis was essential to providing an ontology output with a hierarchical structure to enable easier application in an ICT system, and to be able to form ontology output from different perspectives to suit different applications in different domains. The analysis started from deriving each keyword's social position, namely their centrality in the network.

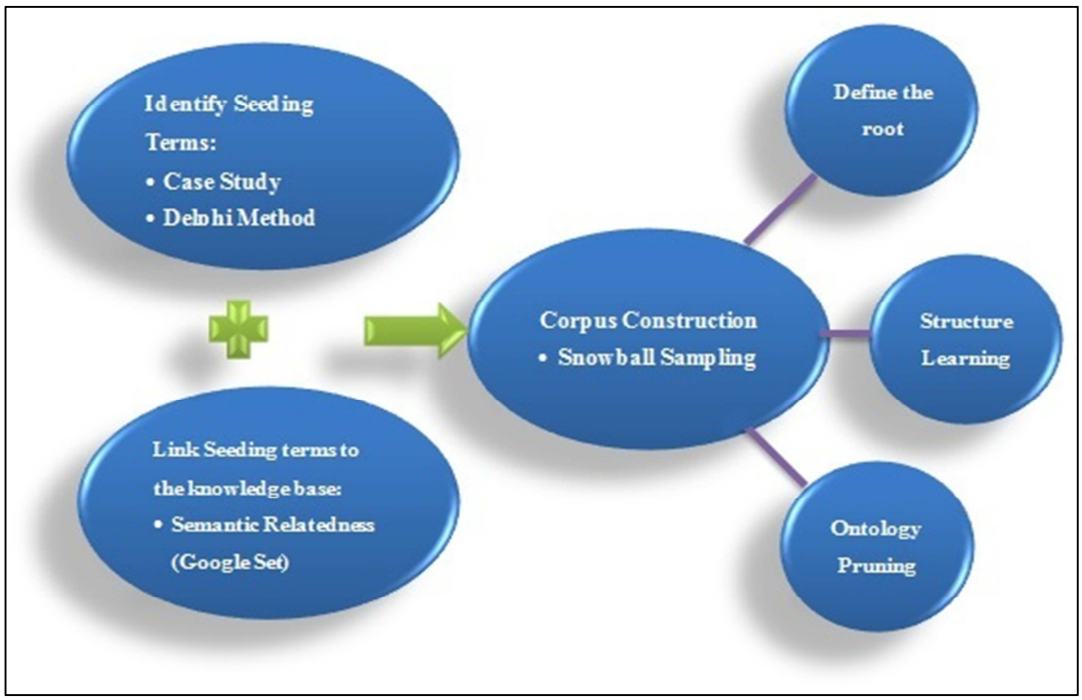

Figure 5: Detailed Techniques for Linking Seeding Words to the Knowledge Base

4.2.1 Centrality Analysis

There were 10660 unique keywords in the prediction sets, and their occurrences varied from once to 3432 times. Those members who had been "derived" (linked by others) more times could be regarded as more representative of the group, or more "centrally" located within a concept. Such

centralised terms are the super connectors among groups of keywords (analogous to key social network members) within the overall network[28].

The corpus construction described in the experiment resulted in n(n-1)/2 sets of collections. To examine the centrality of a target member (*m*) in such a data structure, the calculation had to go through every collection to count the possible relations it has with all the possible seeding words. Thus, the centrality algorithm had two steps:

Firstly, verifying the existence of (*m*) in every collection or Set (*S*), under the conditions that Set (*S*) was not seeded by a pair of words including (*m*) itself. The existence of (*m*) in Set (*S*) was configured as $f_E(m,S)$ to generate a numeric value.

$$f_E(m,S) = \begin{cases} 1, m \in S \\ 0, m \notin S \end{cases} \mid f_{GS}(m,k) \neq S \right\}$$
$$Where: \quad S = \forall S_{(k_{pi},k_{pj})} \mid 1 \leq i < j \leq n$$
$$And \quad m \in \{k_{p1}, k_{p2}, \ldots, k_{pn}\},$$
$$k \in \{k_{p1}, k_{p2}, \ldots, k_{pn}\}, and \quad m \neq k$$

Then, the total connections of (*m*) in these sets are the aggregation of $f_E(m,S)$. This can be calculated as the centrality:

$$f_{Cn}(m) = \sum_{i,j} f_E(m, S_{(k_{pi},k_{pj})}) \mid 1 \leq i < j \leq n$$

Among 10660 generated keywords, 3920 keywords only appeared once. A one-time appearance implies that the predicted word does not have close connections with the other keywords but remotely connects with only one pair. For the purposes of this research, we define “one time appearance” as noise in the experiment. The remaining keywords are distributed as shown:

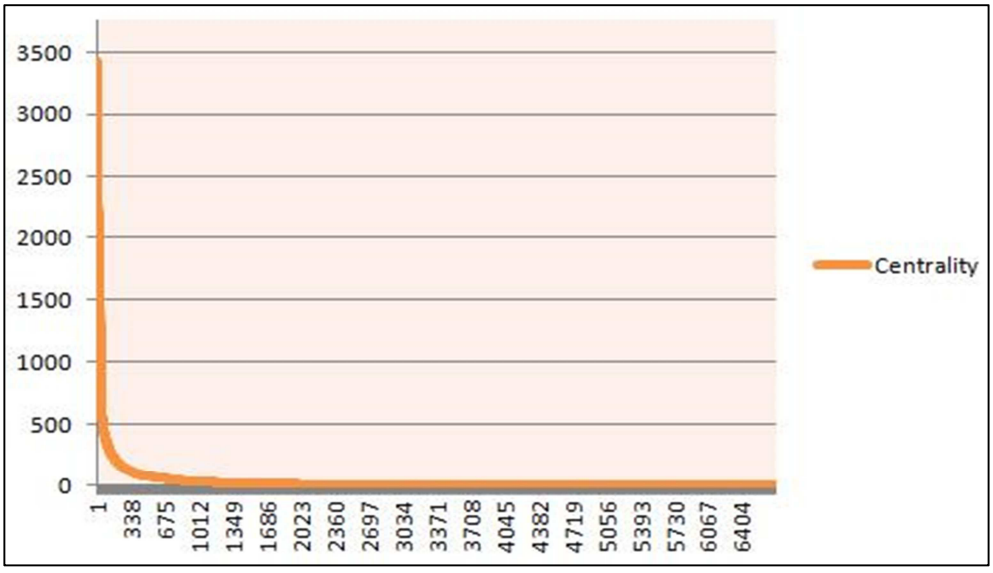

Figure 6: Keywords Centrality Analysis

The distribution in figure 6 is similar to a Poisson distribution. To understand more about the curve, we could cut it into three pieces by tangent (**y = -x)**. Then the curve would be divided into three core zones (Figure 7):

1. Curve 1 (definition zone) presents a fully connected top zone with highly centralised members. Mathematically, these keywords appeared much more often than the other members outside the zone
2. Curve 2 (description zone) shows a fast drop that indicates those keywords used quite often as descriptors in the domain. Their centralities were lower than the top definition zone, but most of them were connected to top zone members.
3. Curve 3 (connection zone) includes those low centralised keywords mentioned around the concept, but not necessarily a part of the concept, although they do have some connection with some of the words in the definition or description zone.

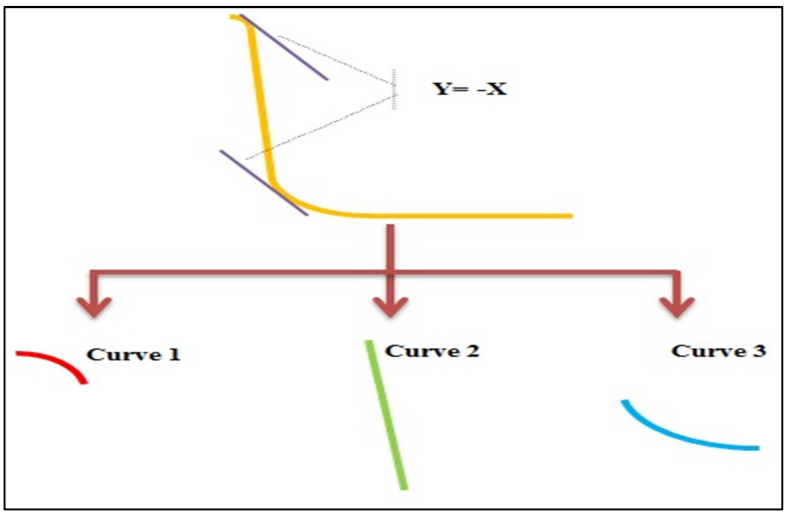

Figure 7: Cut-off Points

4.2.2 Closeness Analysis

"Closeness" analysis helped to shape the conceptual clusters around the centralised concepts to provide a more comprehensive description of the concepts and clarify the relationships among them. "Closeness" analysis takes concepts within a domain as observation objects to measure how close concepts are to each other. Unlike centrality analysis, it counts the connections to a concept from another concept. Closeness could be treated as the relevant connective power between concepts. This relevant power can indicate the "closeness" between concepts. In addition, the sum of connections provided a numeric value, and it could be converted (a simple method is to use reciprocal) to a value from 0-1, which could represent the distance between conceptual clusters.

In this research, the closeness investigated how important a seeding word (*k*) was in predicting (*m*), and in semantic relatedness terms, how much did seeding word (*k*) determine the appearance of prediction (*m*) in the domain. Centrality analysis defined $f_{Cn}(m)$ to track (*m*) appearances in all the prediction sets, regardless of their seeding words. If seeding words were considered, for example a seeding word *k*, $f_{cl}(m,k)$ can calculate *m*'s appearances via a traversal of these sets, based on *k*.

$$f_{cl}(m,k) = \sum_{i=1}^{n} f_E(m, S_{(k_{pi},k)})$$

Then, the decisive power of seeding word *k* on predictions *m* could be presented as a closeness distance $f_d(m,k)$. The greater $f_d(m,k)$ is, the greater the decisive power *k* has to predict *m*.

$$f_d(m,k) = \frac{f_{cl}(m,k)}{f_{Cn}(m)}$$

The result of practical closeness analysis on the corpus confirmed that different seeding words had different decisive powers over the number of appearances of a target word. A quantified value helped to refine the zone definition from centrality analysis, as centrality analysis can only conduct zone specification from a structure perspective.

The new methodology generates connections between different terms that are weight specified directional relationships (like vectors) based on the "closeness" value. Such relationship expresses the binary relationship more richly than simple weightless connection. For example, table 1 demonstrates the relationship between several terms to the concept "turning".

Seeding Words (k)	**Predict(m)**	**$f_{cl}(m,k)$**	**$f_{Cn}(m)$**	**$f_d(m,k)$**	**Relevant Distance**
Reaming	Turning	115	2664	0.043168	1
Tapping	Turning	106	2664	0.039790	1.084906
Threading	Turning	97	2664	0.036411	1.185567
Conventional	Turning	93	2664	0.034910	1.236559
Screw cutting	Turning	93	2664	0.034910	1.236559
Drilling	Turning	79	2664	0.029655	1.455696
Centering	Turning	79	2664	0.029655	1.455696
Micro drilling	Turning	72	2664	0.027027	1.597222
Deburring	Turning	67	2664	0.025150	1.716418
Cutting	Turning	65	2664	0.024399	1.769231
CNC Machining	Turning	26	2664	0.009760	4.423077
Thread rolling	Turning	22	2664	0.008258	5.227273

Table 1: Weight Specified Relationship

Drilling and Centering can be associated with either Turning or Milling. The "distracted" linkage towards both Turning and Milling may reduce the strength of the relationships towards either of them. Therefore, they appeared to be "less strongly" related to turning process.

4.2.3 Betweenness Analysis

"Betweenness" analysis was implemented to assist in uncovering the overall structure of the network to identify the bridging elements that connect every member together in the domain network structure. It identifies those members whose importance may be missed by centrality and

closeness analysis but who bridge the gaps between concept clusters. Betweenness analysis finds those individuals or groups who have concurrent membership in overlapping concepts, so the relations between concepts become clearer. In this research, members with significant "Betweenness" factors were found via the following method:

1. Reference to the closeness addressed those members with a low closeness in the network; this meant that such concept clusters were semantically further apart than others. In this research, special attention was paid to those members that are remotely positioned in both directions. For instance, the traversal of f_d could address predictions m_1 and m_2, where:

$$f_d(m_1, m_2) \to 0 \quad \text{and} \quad f_d(m_2, m_1) \to 0$$

Addressing this sort of relationship was the key to clarifying the conceptual clusters, especially when both m1 and m2 were highly centralised members. It provided numerical figures to draw a boundary between m1 and m2.

2. But there may exist a prediction k which is decisive for both m1 and m2:

$$f_d(k, m_1) \to \max_{1 \le i \le n} f_d(k, k_{pi})$$

$$\text{and} \quad f_d(k, m_2) \to \max_{1 \le i \le n} f_d(k, k_{pi})$$

Such *k* connected m_1 and m_2 from *k*'s view point. The existence of such a keyword shows that a bridging concept exists and could be located. It also indicates that the peripheral players of a network should not be omitted, since they may be the bridge to other networks.

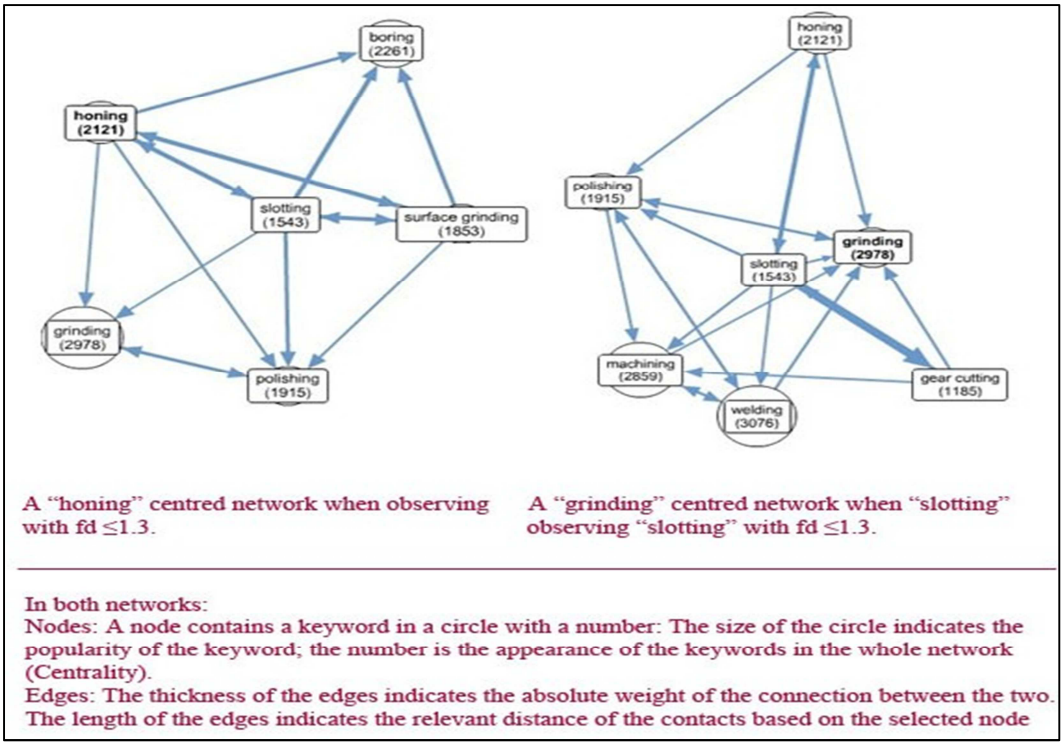

Figure 8: Illustration of the Engineering Ontology Network

Seeding Words (k)	Predict (m)	$f_{cl}(m,k)$	$f_{Cn}(m)$	$f_d(m,k)$
Folding	Honing	3	2121	0.001414
Honing	Folding	1	1131	0.000884
Tool grinding	Honing	83	2121	0.039132
Tool grinding	Folding	58	1131	0.051282

Table 2: Example of the Betweenness Analysis in the Engineering Ontology

The analysis revealed that this method of analysis was able to create well positioned "betweenness" measures between members. For example, table 2 shows that "folding" and "honing" in the generated engineering ontology are not particularly close to each other. However, there was a member "tool grinding" which is tightly connected to both of them. (Figure 8)

5 Discussion of the result

The research also investigated that the process is repeatable, that cut-off points were set reasonably, that the final output served the research objectives, and that the research could be applied to real life environments.

5.1 Zones Explanation

5.1.1 Connection zone

The ground level connection zone contains "long tail" terms nominated by the terms in the two upper levels. Terms in the ground level did not necessarily describe the main concepts accurately, but they were connected to the concepts or concepts' descriptions to some extent in the domain context. For example, "food processing" was identified as a connection zone member in the new engineering ontology. Practically, such a connection zone member does have a relationship with the main concepts. However, the frequency of appearance of the terms in this zone was the lowest in the three zones. These third zone terms were valuable from other perspectives: in terms of structural clarification such members could be boundary players and from a cross domain viewpoint they may be the brokers from the target domains to related domains.

5.1.2 Description Zone

This zone comprises popular concepts or terms defining in more detail the concepts from the top zone. Observation of these concepts or terms revealed that many of them were phrases containing concepts or their synonyms from the top definition zone. At this level, terms were inevitably connected to the relevant concepts at the top level but were not as important as them (lower centrality value). For example, "drilling" is a core concept in the new engineering ontology; its directly linked concepts "gun drilling" and "cross drilling" are description zone members.

Members in the description zone have at least one direct connection to a few but not all of the top zone members, and additionally they have limited connections with each other. Not being able to form a complete network is a distinguishing characteristic of the remainder of the network members. An incomplete network also implies separation of their corresponding concepts (or

conceptual clusters), thus borders could be drawn based on such disconnectivity. Although not fully connected, these members can reach all top level members and most of the other descriptive members within three steps as required by network reach analysis.

5.1.3 Definition Zone

Compared with other zones, the keywords in the definition zone appear more often, and they are thus the keywords that define the domain(s) most explicitly.

Keyword	Centrality	Keyword	Centrality
Drilling	3432	Centering	1862
Welding	3330	Conventiona	1852
Milling	3157	Slotting	1776
Machining	3148	Electroformin	1747
Grinding	3128	Screw	1741
Cutting	3012	Tool	1667
Tapping	2879	Gear shaping	1660
Sawing	2824	Stamping	1644
Turning	2789	Micro	1643
Painting	2771	Finishing	1511
Assembly	2765	Fabrication	1490
Punching	2685	Gear cutting	1482
Bending	2468	CNC	1456
Boring	2408	Rolling	1263
Deburring	2344	Heat treating	1216
Forming	2331	Laser cutting	1206
Honing	2305	Folding	1169
Broaching	2270	Plating	1106
Shearing	2192	Notching	1095
Polishing	2144	Custom	1002
Threading	2125	Engineering	919
Reaming	2080	Powder	912
Surface	2077	Design	912
Cylindrical	1919	Thread	901
Surfacing	1896	Plasma	856

Table 3: Definition Zone Members

In the definition zone, members cover most of the WMCCM categories and the UK SIC codes for the engineering area. For example, [6] describes machining (first column in Table 3) as:

"*This class includes:*

- *cutting, boring, turning, milling, eroding, planing, lapping, broaching, levelling, sawing, grinding, sharpening, polishing, welding, splicing etc. of metalwork pieces*

- *cutting of and writing on metals by means of laser beams.***"**

Nine out of fifteen keywords in the SIC definition are covered by the definition zone, with the remainder covered by the lower zones (4 by the description zone and 2 by the connection zone). In addition, the research generates all the WMCCM categories that exist in the set. WMCCM proposed 22 concepts in the definition zone (second column in Table 4). With the new ontology, 16 out of 22 of these concepts were covered by the definition zone and another three have high centrality in the description zone, with the rest covered by the connection zone. Moreover, the prediction set generated covers more domain space than both the SIC and WMCCM ontology. The results provide evidence that they are not only accurate, but also have a wider coverage than the standard code (Table 4).

SIC	WMCCM Ontology	New Ontology	Centrality
Boring	Boring	Boring	2408
Broaching	Broaching	Broaching	2270
	CNC Laser Cutting	Laser Cutting	1206
	CNC Machining	CNC Machining	1456
	CNC Milling	CNC Milling	511
	CNC Turning	CNC Turning	405
Cutting	Cutting	Cutting	3012
	Drilling	Drilling	3432
Eroding		Eroding	64
	Fettling	Fettling	2
	Gear Cutting	Gear Cutting	1482
Grinding	Grinding	Grinding	3128
	Hobbing	Hobbing	2305
	Manual Machining	Machining	3148
Lapping		Lapping	289
Levelling		Levelling	25
Milling	Milling	Milling	3157
Planning		Planning	58
Polishing		Polishing	2144
	Profiling	Profiling	143
Sawing	Sawing	Sawing	2824
	Splining	Splining	37
Sharpening		Sharpening	92
Splicing		Splicing	2
	Tapping	Tapping	2879
	Thread Grinding	Thread Grinding	42
	Threading	Threading	2125
Turning	Turning	Turning	2789
Welding	Welding	Welding	3330

Table 4: Ontology Content Comparison

5.2 Repeatability

Similar experimentation has also been conducted for the other domains to assess if the appearance curve will remain the same shape. This showed the same trend as engineering: a fairly short definition zone, a sharp drop description zone and a very long tail connection zone. Such repetition of the curves indicated that the predictions do maintain the same trend and the experiment is repeatable.

5.3 Fault tolerance

Another valuable contribution of the research is that the process has a fault tolerance ability. Originally, the research was designed to have three pairs of keywords to avoid potential misdirection by a badly chosen term. Three pairs will allow one pair to be misleading, but will still have 66.7% outputs towards the right direction in theory.

In fact, we did have a bad sample in our experiment: one of our original chosen words was "hobbing", and its appearance was only 120, which made it fall into the connection zone. But contrarily, this expresses the fault tolerance ability of the system: 'hobbing' is recognised in the connection zone, so it has a quite limited effect on the other two more important zones.

5.4 Optimisation of the WMCCM current process

The derived ontology can be applied to optimise an existing ontology (via different integration methods such as ontology merging or alignment [29]): if the existing ontology requires its basic concepts to remain unchanged, the new ontology can be centralised on those concepts. Alternatively, the derived methodology could enrich the existing ontology structure by adding descriptive concepts and relationships found by this approach. An implementation of the derived ontology was evaluated to solve practical problems in information categorisation for WMCCM - the ontology is formed according to the zones' definition and is practically used for the subsequent categorisation of the tenders. A monitoring mechanism was implemented to compare the performance of the original engineering ontology used by WMCCM and the ontology developed through this research. 5101 engineering tenders were processed through the system. Figure 9 demonstrates that the categorisation system has been improved by adopting the new ontology:

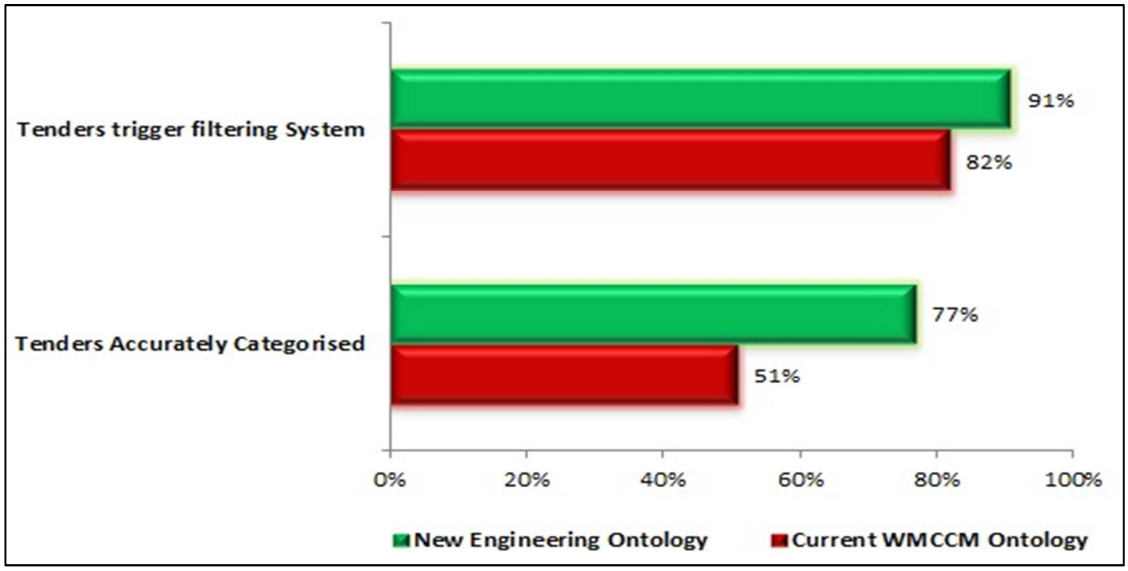

Figure 9: Practical Evaluation of the New Engineering Ontology

- The new ontology filter was triggered by 91% of the input information (4640), compared to 82% (4193) with the existing WMCCM ontology.
- Among those filtered items, 77% (3569) of the information had appropriate categorisation with the new ontology, compared to only 51% (2152) being correctly categorised by the existing one. This was due to insufficient internal relationships within the existing ontology.

Practical evaluation proved that the new derived ontology can be fitted to the desired automated system and provided statistically better categorisation results. Moreover, the new ontology could be fitted to an existing fixed ontology by adding the generated rich concepts and relationships as conceptual descriptions (Such descriptions only supplement additional terms and relationships without changing the ontological structure).

6 Conclusion

Good ontology can play a key role in information systems for “intelligent” processing and categorisation. Through the investigation of the WMCCM ontology and other relevant ontology in the engineering and manufacturing domain, the need was identified to quickly, reliably and economically generate ontology that are able to provide the breadth and depth of coverage required for the given domain. This is particularly important to building multidisciplinary or cross domain systems.

A new ontology development methodology has been proposed to address those needs, and the derived ontology for an engineering case study has been implemented and evaluated. The derived ontology addresses the issues regarding the cost of generating ontology with sufficient scope and relationships richness. It has been demonstrated that a rich multi-disciplinary ontology can be built with only three pairs of seeding words provided by a domain expert using a semantic-relatedness-based tool. This ontology has a high breadth and depth of concept coverage and derives internal relationships to form a network structure. The evaluation of the derived ontology has demonstrated that it has performed better in the automated information categorisation applications than the current ontology adopted by WMCCM. This technique has huge potential in automating the handling of human queries, through better interpretation and categorisation abilities realised through much richer semantic relationships and broader domain coverage.

References

[1] P. Timmers, Business Models for Electronic Markets, Electronic Markets, 8 (2). (1998). 3-8.

[2] Y. Cheung, H. Scheepers, M. Swift, V. Lee, J. Bal, A Competence-Based Collaborative Network: The West Midlands Collaborative Commerce Marketplace, in: L. Camarinha-Matos, X. Boucher, H. Afsarmanesh (Eds.), Collaborative Networks for a Sustainable World, Springer Berlin Heidelberg, 2010, pp. 380-387.

[3] G. van Heijst, A.T. Schreiber, B.J. Wielinga, Using explicit ontologies in KBS development, International Journal of Human-Computer Studies, 45 (1). (1997). 183-292.

[4] R. Mizoguchi, J. Van Welkenhuysen, M. Ikeda, Task Ontology for Reuse of Problem Solving Knowledge, in: N.J.I. Mars (Ed.), Towards Very Large Knowledge Bases, IOS Press, Amsterdam, 1995, pp. 60-72.

[5] M. Jarrar, R. Meersman, Ontology Engineering - The DOGMA Approach, in: S.D. Tharam, C. Elizabeth, M. Robert, S. Katia (Eds.), Advances in Web Semantics I, Springer-Verlag, 2009, pp. 7-34.

[6] L. Prosser, UK Standard Industrial Classification of Economic Activities (SIC 2007), Available: www.ons.gov.uk/ons/guide-method/classifications/current-standard-classifications/standard-industrial-classification/sic2007---explanatory-notes.pdf. [Accessed on 27/08/2009]

[7] O. Corcho, A. Gómez-Pérez, Solving Integration Problems of Ecommerce Standards and Initiatives through Ontological Mappings, in: A. Gómez-Pérez, M. Grüninger, H. Stuckenschmidt, M. Uschold (Eds.), IJCAI'01 Workshop on Ontologies and Information Sharing, CEUR-WS.org, Seattle, Washington, 2001, pp. 131-140.

[8] D.B. Lenat, R.V. Guha, Building Large Knowledge-Based Systems; Representation and Inference in the Cyc Project, Addison-Wesley Longman Publishing Co., Inc., 1989.

[9] M. Grüninger, M.S. Fox, Methodology for the Design and Evaluation of Ontologies, in, Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing (IJCAI95), Conference, 1995.

[10] S. Staab, H.-P. Schnurr, R. Studer, Y. Sure, Knowledge processes and ontologies, IEEE Intelligent Systems, 16 (1). (2001). 26-34.

[11] M. Fernández-López, A. Gómez-Pérez, N. Juristo, METHONTOLOGY: from Ontological Art towards Ontological Engineering, in, Proceedings of the AAAI-97 Spring Symposium Series, Stanford University, EEUU, Conference, American Asociation for Artificial Intelligence, 1997.

[12] B. Swartout, P. Ramesh, K. Knight, T. Russ, Toward Distributed Use of Large-Scale Ontologies, in: G.M. Farquhar A, Gómez-Pérez A, Uschold M, van der Vet P (Ed.), AAAI'97 Spring Symposium on Ontological Engineering, Stanford University, California, 1997, pp. 138-148.

[13] M. Fernández-López, A. Gómez-Pérez, J.P. Sierra, Building a Chemical Ontology Using Methontology and the Ontology Design Environment, IEEE Intelligent Systems, 14 (1). (1999). 37-46.

[14] A. Bernaras, I. Laresgoiti, J.M. Corera, Building and Reusing Ontologies for Electrical Network Applications, in, Proceedings of the 12th European Conference on Artificial Intelligence, Budapest, Hungary, Conference, 1996.

[15] A. Gómez-Pérez, Knowledge Sharing and Reuse, in: J. Liebowitz (Ed.), Handbook of Expert Systems, CRC Press, Boca Raton, Florida, 1998.

[16] R.Y.K. Lau, Fuzzy Domain Ontology Discovery for Business Knowledge Management, The IEEE Intelligent Informatics Bulletin, 8 (1). (2007). 29-41.

[17] E. Agirre, O. Ansa, E. Hovy, D. Martínez, Enriching very large ontologies using the WWW, in: S. Staab, A. Maedche, C. Nedellec, P. Wiemer-Hastings (Eds.), Proceedings of the the First Workshop on Ontology Learning OL'2000, Berlin, Germany, Conference, 2000.

[18] L. Qi, H. Daqing, M. Ming, A Study of Relation Annotation in Business Environments Using Web Mining, in, Proceedings of the IEEE International Conference on Semantic Computing, 2009. ICSC '09., Conference 14-16 Sept. 2009, 2009, pp. 203-208.

[19] M.d. Kunder, Size of The World Wide Web, Available: http://www.worldwidewebsize.com. [Accessed on 08/08/2012]

[20] A. Gulli, A. Signorini, The indexable web is more than 11.5 billion pages, in, Proceedings of the Special interest tracks and posters of the 14th international conference on World Wide Web, Chiba, Japan, Conference, ACM, 2005, pp. 902-903.
[21] S. Tong, J. Dean, System and methods for automatically creating lists, in, USA, 2008.
[22] H.A. Linstone, M. Turoff, The Delphi Method: Techniques and Applications, Addison-Wesley, Reading, Mass., 1975.
[23] M.J. Salganik, D.D. Heckathorn, Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling, Sociological Methodology, 34. (2004). 193-239.
[24] O. Frank, Network Sampling and Model Fitting, Cambridge University Press, New York, 2005.
[25] M. Swift, N. Armoutis, J. Bal, M. Molfetas, The Formation of Virtual Organisations to Address Complex Tenders through a Collaborative Commerce Marketplace, in, Proceedings of the The 13th International Conference on Concurrent Enterprising, Sophia-Antipolis, France, Conference, 2007.
[26] A.M. Fairchild, B. De Vuyst, Coding Standards Benefiting Product and Service Information in E-Commerce, in, Proceedings of the The 35th Hawaii International Conference on System Sciences, Hawaii, Conference, IEEE Computer Society, 2002.
[27] E.K. Jacob, Classification and Categorization: A Difference that Makes a Difference, LIBRARY TRENDS, 52 (3). (2004). 515–540.
[28] L. Katz, A new status index derived from sociometric analysis, Psychometrika, 18 (1). (1953). 39-43.
[29] M. Klein, Combining and relating ontologies: an analysis of problems and solutions, in, Proceedings of the Conference Name|, Conference Location|, Conference Date|, Publisher|, Year of Conference|, pp. Pages|.

Appendix A - Detailed Algorithms for Snowball Sampling

As described in section 3.4 the corpus construction methodology may be represented by the following formula.

$$S_{(x,y)} = f_{GS}(x,y) = \left\{k_1^{x,y}, k_2^{x,y}, \dots, k_{n_{x,y-1}}^{x,y}, k_{n_{x,y}}^{x,y}\right\}$$

Function $f_{GS}(x,y)$ is the process to capture Google Sets results by using given paired seeding keywords x and y. Set $S_{(x,y)}$ is the collection of keywords from $k_1^{x,y}$ to $k_{n_{x,y}}^{x,y}$ generated from seeding keyword pair x and y.

The multiple iterations of this formula in achieving the "snowballing" effect are explored in more detail here.

In the applied methodology, k1&k2, k3&k4, k5&k6 are defined as three pairs of keywords selected for a chosen domain/application *M* (where *M* is the concept/definition of the domain(s)). These keywords are usually supplied by domain experts, or maybe taken from an existing ontology.

Function $f_{GS}(x,y)$ is the process to capture Google Sets results by using given paired seeding keywords x and y. Set $S_{(x,y)}$ represents the collection of the predicted keywords, from $k_1^{x,y}$ to $k_n^{x,y}$ which were generated by function $f_{GS}(x,y)$.

$$S_{(k1,k2)} = f_{GS}(k1,k2) = \left\{k_1^{1,2}, k_2^{1,2}, \dots, k_{(n_{1,2}-1)}^{1,2}, k_{n_{1,2}}^{1,2}\right\}$$

Then, in order to generate more optimised outputs, the second round collects the predictions from the first round and pairs them up with the original seeding words as new seeding pairs, and then obtains the new extended predictions from Google Sets. Extended collection for k_1&k_2:

$$S_{\left(k_1,k_1^{1,2}\right)} = f_{GS}\left(k_1,k_1^{1,2}\right) = \left\{k_1^{1,1,1,2}, k_2^{1,1,1,2}, \dots, k_{(n_{1,1,1,2}-1)}^{1,1,1,2}, k_{n_{1,1,1,2}}^{1,1,1,2}\right\}$$

$$\vdots$$

$$n_{1,2}$$

$$\vdots$$

$$S_{\left(k_1,k_{n_{1,2}}^{1,2}\right)} = f_{GS}\left(k_1,k_{n_{1,2}}^{1,2}\right) = \left\{k_1^{1,(n_{1,2}),1,2}, k_2^{1,(n_{1,2}),1,2}, \dots, k_{(n_{1,(n_{1,2}),1,2}-1)}^{1,(n_{1,2}),1,2}, k_{n_{1,(n_{1,2}),1,2}}^{1,(n_{1,2}),1,2}\right\}$$

$$S_{\left(k_2,k_1^{1,2}\right)} = f_{GS}\left(k_2, k_1^{1,2}\right) = \left\{k_1^{2,1,1,2}, k_2^{2,1,1,2}, \ldots, k_{(n_{2,1,1,2}-1)}^{2,1,1,2}, k_{n_{2,1,1,2}}^{2,1,1,2}\right\}$$

$$\vdots$$

$$n_{1,2}$$

$$\vdots$$

$$S_{\left(k_2,k_{n_{1,2}}^{1,2}\right)} = f_{GS}\left(k_2, k_{n_{1,2}}^{1,2}\right) = \left\{k_1^{2,(n_{1,2}),1,2}, k_2^{2,(n_{1,2}),1,2}, \ldots, k_{(n_{2,(n_{1,2}),1,2}-1)}^{2,(n_{1,2}),1,2}, k_{n_{2,(n_{1,2}),1,2}}^{2,(n_{1,2}),1,2}\right\}$$

The same formula is applied to the rest of the first round predictions. Then "snowballing" to get a wide domain coverage, all the unique predictions from the second round (from k_{p1} to k_{pn}) were re-paired to become the seeding pairs of the third round to generate the final keyword predictions. In theory this process could be repeated until no unique predictions remained, but in practice we found three rounds were sufficient for most domains. The breadth of the predictions is determined by the number of seeding words and the depth by the number of rounds of snowballing. If there are (n) unique predictions from the second round, then the seeding word pairing possibility would be n(n-1)/2, according to the previous formulas.

$$S_{(k_{p1},k_{p2})} = f_{GS}(k_{p1}, k_{p2}) = \left\{k_1^{p1,p2}, k_2^{p1,p2}, \ldots, k_{(n_{p1,p2}-1)}^{p1,p2}, k_{n_{p1,p2}}^{p1,p2}\right\}$$

$$\vdots$$

$$S_{(k_{p(n-1)},k_{pn})} = f_{GS}(k_{p(n-1)}, k_{pn}) = \left\{k_1^{p(n-1),pn}, k_2^{p(n-1),pn}, \ldots, k_{(n_{p(n-1),pn}-1)}^{p(n-1),pn}, k_{n_{p(n-1),pn}}^{p(n-1),pn}\right\}$$

Thus we have described a novel methodology which generates ontology for a specified domain(s) economically, quickly and reliably, resulting in a well-populated set of semantically related domain terms.